भारतीय मानक
Indian Standard

IS 16716 : 2018

स्थान और पैमाने पैरामीटर के लिए मजबूत आकलन

Robust Estimation for Location and Scale Parameters

ICS 03.120.30

© BIS 2018

भारतीय मानक ब्यूरो
BUREAU OF INDIAN STANDARDS
मानक भवन, 9 बहादुरशाह ज़फर मार्ग, नई दिल्ली–110002
MANAK BHAVAN, 9 BAHADUR SHAH ZAFAR MARG
NEW DELHI-110002
www.bis.org.in www.standardsbis.in

June 2018 **Price Group 4**

Statistical Methods for Quality and Reliability Sectional Committee (MSD 03)

FOREWORD

This Indian Standard was adopted by the Bureau of Indian Standards, after the draft finalized by Statistical Methods for Quality and Reliability Sectional Committee had been approved by Management and Systems Division Council.

Many a times a sample data may give misleading results owing to presence of outliers. Presence of outliers also causes non-normality of sample distribution. Even a small proportion of outliers can produce a distorting result. Classical estimators like location parameter and scale parameter, optimal under normality assumption are extremely sensitive to presence of a small fraction of outliers.

One way to handle outliers is to detect them by suitable methods and discard them before any conclusion is drawn from the sample. However, this involves losing some data points which may contain some valuable information. As an alternative, robust method of estimation has been advocated. Robust in the sense, that the outliers do not have any or little impact on the estimation of parameters through these methods.

Robust statistics is a convenient modern way of summarizing results when we suspect that they include a small proportion of outliers. Most estimates of central tendency (*for example*, the arithmetic mean T_n) and of dispersion (for example, standard deviation S_n) depend for their interpretation on an implicit assumption that the data comprise a random sample from a normal distribution. But we know that analytical data often depart from that model. They are often heavy tailed (contain a higher than expected proportion of results far from the mean) and sometimes contain outliers.

Let's consider an example data set: 4.5, 4.9, 5.6, 4.2, 6.2, 5.2, and 9.9. The value 9.9 is clearly a suspect, even in such a small sample. If we include the suspect value in the calculations, we obtain: $T_n = 5.8$, $S_n = 1.9$. These statistics, used to define a model based on a normal distribution, describe the data, but not well. The mean seems to have a high bias, while the standard deviation seems too large. Moreover, the numerical values of these estimates, especially that of the standard deviation, are highly dependent on the actual value of the suspect value.

A more reasonable interpretation of these data is that they comprise a random sample from a population with a mean of about 5 and a standard deviation of about 1, with an outlier at 9.9. If we exclude the outlier from the calculations we find $T_n = 5.1$, $S_n = 0.7$. These statistics provide a plausible normal model for most of the data. Although it provides us with no warning about the possible presence of outliers, this model is often preferable in applications in analytical science.

Indian Standard

ROBUST ESTIMATION FOR LOCATION AND SCALE PARAMETERS

1 SCOPE

In this standard, robust method of estimation of location and scale parameter has been described. Some criteria for measuring 'Robustness' have also been discussed. Among robust estimation procedures, M-estimation has been described in detail. A numerical example of M-estimation is provided. Two other robust estimators, namely, 'Trimmed Mean' and 'Winsorized Mean' are also described. M-estimator, Inter-quartile range, Median Absolute Deviation (MAD) and Gini's Mean Difference are the robust scale estimators that are discussed in this document.

2 TERMINOLOGY

For the purpose of this standard the following definitions shall apply.

2.1 Outliers — No observation can be guaranteed to be a totally dependable manifestation of the phenomena under study. The probable reliability of an observation is reflected by its relationship to other observations that were obtained under similar conditions. Observations that in the opinion of the investigator stand apart from the bulk of the data are called outliers. An outlier is one that appears to deviate markedly from the other members of the sample in which it occurs.

2.2 Robust Estimation — Robust statistics provides an alternative approach to classical statistical methods. The motivation is to produce estimators that are not excessively affected by small departures from model assumptions. These departures may include departures from an assumed sample distribution or data departing from the rest of the data (that is outliers).

2.3 Robustness Criteria

2.3.1 *Sensitivity Curve* **—** The sensitivity curve of the estimate $\hat{\mu}$ (say) for the sample $x_1,\ldots,x_n$ is the difference:

$$\hat{\mu}(x_1,\ldots,x_n;x_0)-\hat{\mu}(x_1,\ldots,x_n)$$

as a function of the location x_0 of an outlier. This is to see what happens to an estimator when we change the distribution of the data slightly: it assumes a distribution, and measures sensitivity to change in this distribution.

2.3.2 *Influence Function (IF)* — The influence function is a local measure of robustness, representing the effect of a single outlier. In contrast to sensitivity curve, the empirical influence assumes a sample set, and measures sensitivity to change in the samples. The influence function of an estimator is an asymptotic version of its sensitivity curve. It is an approximation to the behavior of an estimate when a sample contains a small fraction of identical outliers.

2.3.2.1 *Example* — Consider the sample x_1, x,.., x_n and assume that the observations are ordered $x_1 \leq x_2 \leq \ldots \leq x_\mathrm{n}$ with sample size *n* to be odd. The influence function measures the effect of small changes in the distribution (the sample) on the value of an estimator. The word 'small' is a subjective and rather vague notion. It has to be made precise in each particular problem that is studied. Assume that one replaces the observation x_n above by the observation $x_\mathrm{n} + \varsigma$ This change causes a deviation from the original *i.i.d.* assumption. The change is small in that only one out of the *n* original observations is replaced. In a quite different sense, however, the change is substantial, as ς may diverge to infinity. So the number of changed observations (that is, the fraction of contamination) is small, but the actual change in the contaminated observation may be excessively large.

For example, consider the mean, which is given by: $\hat{\mu}=\frac{1}{n}\sum_{t=1}^{n}x_t$ and the median, which is given by $\hat{\mu}=x_{(n+1)/2}$. One can now ask the question how the mean and the median are changed if one of the observations x_n is changed to $x_\mathrm{n}+\varsigma$. Let $\hat{\mu}^{\varsigma}$ denote the estimator for the revised sample. Note that $\hat{\mu}^{0}$ produces the expressions for the mean and median for the original sample. One obtains $\hat{\mu}^{\varsigma}=\hat{\mu}^{0}+\varsigma/n$ for the mean, and

$$\hat{\mu}^{\varsigma}=\begin{cases}\hat{\mu}^{0} & for\,\varsigma > x_{(n+1)/2}-x_n\\ x_n+\varsigma & for\,x_{(n-1)/2}-x_n\leq\varsigma\leq x_{(n+1)/2}-x_\mathrm{n}\\ x_{(n-1)/2} & Otherwise\end{cases}$$

for the median. If one considers the difference between $\hat{\mu}^{\varsigma}$ and $\hat{\mu}^{0}$, standardized by the fraction of changed data points, one obtains $n\ (\hat{\mu}^{\varsigma}-\hat{\mu}^{0})$ for the mean, and

$$n(\hat{\mu}^{\varsigma}-\hat{\mu}^{0})=\begin{cases}0 & for\,\varsigma > x_{(n+1)/2}-x_n\\ n(x_n+\varsigma-x_{(n+1)/2}) & for\,x_{(n-1)/2}-x_n\leq\varsigma\leq x_{(n+1)/2}-x_n\\ n(x_{(n-1)/2}-x_{(n+1)/2}) & Otherwise\end{cases}$$

for the median. This standardized difference is closely related to the finite sample versions of the influence function. Note that for the mean, the standardized difference is an unbounded function in ς. Consequently, by choosing a suitable value for ς, the mean can attain any value by changing only one observation. Stated differently, one outlier can have an arbitrarily large impact on the mean. In contrast, the standardized difference for the median is a bounded function of ς. Therefore, one outlier can only have a limited impact on the median as an estimate of the location of the sample. In this way, one can see that the median is more robust than the mean. The influence function (IF) tries to provide the same information as above by looking at the (standardized) effect of (infinitesimally) small perturbations in a distribution on the value of an estimator.

2.3.3 *Gross Error Sensitivity* — The IF measures the effect of an (infinitesimally) small fraction of contamination on the value of the estimator and is, therefore, called a quantitative robustness concept. In fact, one can say that the IF measures the asymptotic (standardized) bias of the estimator caused by contaminating the target distribution F. Using simple approximations, the IF can be used to compute the numerical effect of small fractions of contamination, for example, one outlier, on the estimator. As the boundedness of its IF is an important feature of an estimator, it is natural to look at the supremum of the IF. This leads us to the notion of gross error sensitivity. The gross error sensitivity r is defined as the supremum of the IF with respect to x. The finite values of r are desirable, because they induce that the estimator changes only slightly if a small fraction of contamination is added. It is easy to show that r is infinite for mean but finite for median.

2.3.4 *Breakdown Point* — Breakdown point is the maximum ratio of unsuitable observations (outliers) such that the estimator still gives reliable estimate. In other word, it is the smallest fraction of anomalous data that can cause the estimator to be useless. The breakdown point of an estimator is the largest fraction of the data that can be moved arbitrarily without perturbing the estimator to the boundary of the parameter space: thus the higher the breakdown point, the more robust the estimator against extreme outliers. The smallest possible breakdown point is $1/n$, that is, a single observation can distort the estimator so badly that it is of no practical use. The breakdown point gives the highest proportion of outliers one may have in the data before the estimator goes over all bounds, and is a global measure of robustness.

The asymptotic breakdown point of the median is 50 percent, the highest possible value. On the other hand, the mean has a zero breakdown point. For a scale estimator S_n, breakdown occurs when the estimator tends to zero or to infinity. The asymptotic breakdown point of the standard deviation is 0 percent, but the Mean Absolute Deviation (MAD) estimators have the highest possible breakdown point of 50 percent. Note that the interquartile range (IQR), has a breakdown point of only 25 percent while it has the same IF as the MAD at symmetric distributions.

Note that the breakdown point is not enough to assess the degree of robustness of an estimator. Instead, the gross error sensitivity gives an exact measure of the size of robustness, since it is the supremum of the influence function of an estimator, and it is a measure of the maximum effect an observation can have on an estimator.

2.3.5 *Maximum Bias* — The influence function and the gross-error-sensitivity can be used to quantify the robustness properties of a given estimate. However, these robustness measures are not completely satisfactory because of the following reasons.

The approximation given by the gross-error-sensitivity only applies to vanishingly small fractions of contamination. Real datasets, on the other hand, contain a certain fixed fraction of contamination-outliers which, of course, does not converge to zero.

The gross-error-sensitivity usually fails to capture important features of the estimation procedure including the possible dependence of the estimate on estimated nuisance parameters. Robust estimates, on the other hand, usually depend on estimated nuisance parameters.

The influence function and the gross-error-sensitivity are not 'consistent' robustness measures.

The global robustness measure called maxbias (maximum asymptotic bias) which is now considered the most complete and accurate measure of robustness for a point estimate. Assume that the true distribution function F_0 lies in some family F. Then maxibias of the estimator T is defined over F as

$$b(F) = \mathop{\mathrm{SUP}}_{F^* \in F} \left| T(F^*) - T(F_0) \right| \cdot$$

The family F can have various forms; for example, it can be a neighbourhood of a fixed distribution F_0 with respect to some distance.

3 MERITS OF ROBUST ESTIMATION

It is well known that the widely used least squares estimator is efficient if the errors are independently and identically normally distributed and are independent. When the error distribution is non-normal, the efficiency of the least squares estimator deteriorates. For example, when the error distribution has flatter tails than the normal, the least squares estimator can be inefficient relative to other estimators. Robust estimators are based on minimizing some function of a scale-adjusted error term that gives less weight to large error values. They can be asymptotically more efficient than OLS when the tails of the underlying error term distribution are 'thicker' than the normal.

4 ROBUST ESTIMATION

4.1 Let x_1, x_2,...,x_n be sample observations. It is assumed that these observations are generated through a univariate model as $x_i = \mu + \sigma\,\varepsilon_i$, where ε_i are *i.i.d.* with a common distribution function $F\mu,\sigma$. We assume that this distribution function is normal. The parameter μ is the location parameter, and σ, assumed to be strictly positive, is the scale parameter. In other words, μ is the expected value and the standard deviation of the distribution $F\mu,\sigma$.

4.2 Let T_n be an estimator for the location parameter μ, and S_n an estimator for the scale parameter. The estimator S_n is also called an estimator of spread or of dispersion. A basic requirement is that the location and scale estimators are affine equivariant, meaning that

$T_n\,(ax_1+ b, ax_2+b, ...ax_n +b) = aT_n\,(x_1,x_2,...x_n) + b$; and

$S_n\,(ax_1+ b, ax_2+b, ...ax_n +b) = |a|\,S_n\,(x_1,x_2,...x_n)$ for any real numbers a and b.

4.3 Besides equivariant, an estimator should also be precise, as measured by its statistical efficiency. It is important to realize that the statistical efficiency of an estimator depends on the assumed model distribution. For instance, the sample average has the optimal efficiency of 100 percent at the normal distribution, but at other distributions its efficiency may become very low.

4.4 Robust Estimator for Location Parameter

4.4.1 *M-estimator*

The class of M-estimators contains the maximum likelihood (ML) estimator as a special case. If we assume that the data come from the model distribution $F\mu,\sigma$, in general we define a class of robust estimators that minimize a function ρ of the density function f. Thus we can define $\rho = -\log f$. Then the log-likelihood can be written as:

$$\min \sum_{i=1}^{n} -\log f\,(x_i - \mu) = \min \sum_{i=1}^{n} \rho\,(x_i - \mu)$$

Generally $(x_i - \mu)$ is standardized by a robust scale parameter σ. Thus

$$\min \sum_{i=1}^{n} -\log f\left(\frac{x_i-\mu}{\sigma}\right) = \min \sum_{i=1}^{n} \rho\left(\frac{x_i-\mu}{\sigma}\right).$$

This makes sense for function not corresponding to pdfs. Let $\psi = \rho'$, if this exits. Then we will have the equation for estimating μ as:

$$\sum_{i=1}^{n} \psi\left(\frac{x_i-\mu}{\sigma}\right) = 0 \text{ or } \sum_{i=1}^{n} w_i\left(\frac{x_i-\mu}{\sigma}\right) = 0,$$

where $w_i = \psi\left(\frac{x_i-\mu}{\sigma}\right) \Big/ \left(\frac{x_i-\mu}{\sigma}\right)$.

Thus in M-estimation, for estimating μ different observation is getting different weight. From the above equations, it is clear that an iterative method of solution, updating the weights at each iteration is required. Iteration is started with a known value of μ, say a preliminary estimate like ordinary arithmetic mean. Scale parameter σ is replaced by a suitable robust estimate. It is best to take as a preliminary scale estimator $\hat{\sigma}_0$ the median absolute deviation, or any other highly robust estimator of scale.

Here, ψ is not necessarily the score function, but can be any non-decreasing odd function. M-estimators are affine equivariant, and the estimating equation can easily be solved using the Newton-Raphson algorithm with the sample median as starting value. The influence function of a location M-estimate T at the distribution function F takes the following form:

$$IF(x;T,F) = \frac{\psi(x)}{E_F[\psi'(x)]}$$

We conclude from the above equation that the choice of the ψ-function determines the shape of the *IF*. In particular, a bounded ψ-function yields a robust M-estimator. Moreover, it can be shown that univariate M-estimators of location have a 50 percent breakdown point, as soon as *IF* is bounded. This distinguishes them from trimmed means, which have a bounded influence function but not the highest possible value for the breakdown point.

4.4.1.1 *Examples of M-estimators*

The mean corresponds to $\rho\,(x) = x^2$, and the median to $\rho\,(x) = |x|$. The function,

$$\psi(x) = \begin{cases} x & |x| < c \\ 0 & otherwise \end{cases}$$

corresponds to metric trimming and large outliers have no influence at all. The function,

$$\psi(x) = \begin{cases} -c & x < -c \\ x & |x| < c \\ c & x > c \end{cases}$$

is known as metric Winsorizing and brings in extreme observations to $\mu \pm c$. The corresponding $-\log f$ is:

$$\rho(x) = \begin{cases} x^2 & |x| < c \\ c(2|x| - c) & otherwise \end{cases}$$

This estimator is known as Huber's Estimator. Limit as c → 0 is the median, and $c \to \infty$ as the limit is the mean. The value c = 1.345 gives 95 percent efficiency.

Another class of ψ -functions is obtained from the Tukey's biweight function.

$$\psi(x) = |x| \left[1 - \left(\frac{x}{R} \right)^2 \right]^2 .$$

This implements 'soft trimming'. The value R = 4.685 gives 95 percent efficiency.

4.4.1.2 *Example*

We consider a hypothetical data set to illustrate the M-estimation procedure. Following is the data set. Sample size is 31:

5.1	6.2	6.8	7.8	7.8	7.9	7.9	8.2
8.3	8.4	8.6	8.8	9.0	9.0	10.3	11.7
11.9	12.0	12.2	13.7	14.3	14.5	14.8	16.0
17.5	17.7	18.0	25.0	27.0	37.0	130.0	

The last data point seems to be an outlier. Due to presence of this observation, mean of the sample may be highly distorted. The arithmetic mean of this data is 16.56. We now apply M-estimation procedure. Huber's function has been taken for the purpose with $c = 1.5$. Ordinary estimate of mean, that is, 16.56 has been taken to start the iteration. The standard deviation of the data is 22.13. However, in the formula σ is replaced by a robust estimate of σ, MAD. $\text{MAD} \approx 0.6745\sigma$. The calculated value of MAD is 14.92. In the first iteration M-estimate is calculated as 13.54. In the first iteration first 30 observations get equal weight, that is, 1 and the last observation gets weight equal to 0.196 7. In the next iteration this mean is found to be 13.52 with the weight to the last observation is 0.195 2. In the subsequent iterations not much significant change occurred in the estimation. We, therefore, stopped after the second iteration and concluded that M-estimate by Huber's function is 13.52.

4.4.1.3 *Example*

We consider another hypothetical data set. Following is the data set. Sample size is 26:

57.702	59.296	56.166	55.767	51.722	60.446	60.715	67.447
55.27	59.289	54.027	53.199	41.896	63.264	45.798	58.699
50.086	48.89	62.213	45.625	53.923	55.799	56.741	43.145
60.974	50.706						

In the above data set the 8th data point, that is, 67.447 stands out apart. This data point likely to distort the estimate of mean. There are 26 observations in the sample and mean is calculated as 54.95. Here also we apply Huber's M-estimation procedure. Value of c is taken as 1.5. Ordinary estimate of mean, that is, 54.95 has been taken to start the iteration. The standard deviation of the data is 6.37. and MAD =4.30. In the first iteration M-estimate is calculated as 53.57 with a weight to the 8th observation equal to 0.468. Process is converged by third iteration and final M-estimate is calculated as 53.56.

4.4.2 *Trimmed Mean*

It is the mean of the central 1 – 2α (0 < α < 1) part of the distribution, so [α*n*] largest observations and [α*n*] smallest observations are removed where [α] denotes the nearest integer of *a*. 2α trimmed mean is defined as

$$\bar{x}_{T_\alpha} = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} x_{(i)}$$

where $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are the ordered values of $x_1, x_2, \ldots, x_n$. They are also known as order statistics.

In the above example 10 percent Trimmed mean is calculated as 12.41.

4.4.3 *Winsorized mean*

The 2α Winsorized mean is defined as:

$$\bar{x}_{W,\alpha} = \frac{1}{n}[([n\alpha]+1)x_{([n\alpha]+1)} + \sum_{i=[n\alpha]+2}^{n-[n\alpha]-1} x_{(i)} + ([n\alpha]+1)x_{(n-[n\alpha])}]$$

The winsorized mean is computed after all the $[n\alpha]$ smallest observations are replaced by $x_{([n\alpha]+1)}$, and the $[n\alpha]$ largest observations are replaces by $x_{(n-[n\alpha])}$.

In the above example 10 percent Winsorized mean is calculated as 12.99.

4.5 Robust Estimator of Scale Parameter

The sample standard deviation is a commonly used estimator of the population scale parameter, σ. However, it is sensitive to outliers and may not remain bounded when a single data point is replaced by an arbitrary number. With robust scale estimators, the estimates remain bounded even when a portion of the data points are replaced by arbitrary numbers.

4.5.1 *M-estimator for Scale Parameter*

An M-estimator of scale S_n is defined as the solution of the equation:

$$1/n\sum_{i=1}^{n}\rho\left(\frac{x_i-\hat{\mu}}{\sigma}\right)=k, \qquad 0< k <\rho(\infty)$$

The loss function ρ is even, differentiable, and non-decreasing on the positive numbers. To get a consistent estimator at the model distribution $F_{\mu,\sigma}$, we need to set $k = E_F\ [\rho\ (x)]$. It is most convenient to take for the initial location estimator $\hat{\mu}$ the highly robust sample median. A common choice for the loss function ρ is the biweight function as described earlier. It can be seen that the MAD scale estimator is a limiting case of the biweight M-estimator of scale. M-estimators of scale are equivariant, and the influence function of the corresponding scale functional S_n is given by:

$$IF(x;\ S_n,\ F)=\frac{\rho\ (x)-k}{E_F\left[\rho'\ (x)x\right]}$$

Again, a bounded choice for ρ yields a robust estimator. The asymptotic breakdown point of S_n is given by:

$$\varepsilon^*(S,F_0)=\min\left(\frac{k}{\rho(\infty)},1-\frac{k}{\rho(\infty)}\right)$$

which is strictly positive, but not always equal to 50 percent, even for bounded ρ-functions. The equations for M-estimator of scale can be solved using a fixed-point algorithm. The initial value of σ may be taken any highly robust estimate of it, say MAD. MAD has 50 percent breakdown point. However, first step approximation to this M-estimate gives 50 percent breakdown point. The first step approximation is given by:

$$S_n^1\frac{S_n^0}{nk}\sum_{i=1}^{n}\rho\left(\frac{x_i-\hat{\mu}_0}{S_n^0}\right),$$

where initial value of σ is replaced by S_n^0, an MAD estimate and $\hat{\mu}_0$ is a robust estimate of μ.

This estimate is also called a τ-estimate of scale.

4.5.1.1 *Example*

We consider the same example as given in **4.4.1.2**. Here we obtain the first step approximation, that is, a τ-estimate of σ. The ρ function is taken as Huber's function as given earlier with c = 1.5. Estimate of μ is taken as Huber's M-estimate as obtained earlier, that is,13.52. The initial value of σ is taken as MAD estimate, that is, 14.92. The value of k is taken as 1. However, one can obtain its exact value from $k = E_F\ [\rho\ (x)]$. With its values, the τ-estimate of σ is calculated as 13.25. One can go further steps to obtain an M-estimate of σ.

4.5.2 *Interquartile Range*

IQR is defined as $IQR = Q_3 - Q_1$ where Q_1 and Q_3 are the first and third quartiles respectively. For a normal distribution, the standard deviation σ can be estimated by dividing the interquartile range by 1.34898.

4.5.3 *Median Absolute Deviation (MAD)*

Most popular robust estimator of scale:

MAD = $c\ \ Median_i\left(\left|x_i-Median_j(x_j)\right|\right)$, where the inner *median*, Median$_j$ (x_j) is the median of n observations and the outer median, median is the median of the n absolute values of the deviations about the median. For normal distribution, 1.482 6 × MAD can be used to estimate the standard deviation σ.

4.5.4 *Gini's Mean Difference*

Gini's mean difference is defined as:

$$G=\frac{1}{\binom{n}{2}}\sum_{i<j}\left|x_i-x_j\right|$$

If the observations are from a normal distribution, then $\sqrt{\pi}G/2$ is an unbiased estimator of the standard deviation σ.

5 CASE STUDIES TO DETECT OUTLIERS

5.1 This is a case study about testing a mechanical component, which is continuously under stress while the vehicle is in driving/standstill condition, using a test rig. Road Load data was conducted to device a test recipe for testing this load bearing member on the test rig.

5.2 Six prototypes were subjected to identical cyclic stress test which had ability to generate same failure mode. The data is given below in 'Number of cycles to failure':

Prototype Number	*Cycles to Failure (Life)*
1	929 103
2	934 038
3	938 773
4	540 473
5	889 563
6	846 567

5.3 To an engineering eye the 4th observation appears to have failed at relatively less number of cycles. It is imperative to confirm these using statistical tools. Exploratory graphical tools are discussed below to cross check the hypothesis that this sample is from a common distribution.

5.4 There are several methods using which the outliers can be detected. Three simple and exploratory data based methods are detailed and used to confirm whether the 4th observation is an outlier.

5.4.1 *Exploratory Data Analysis Method: Individual Value Plot*

a) An individual value plot is a graph that can be used to assess sample distributions. It also helps in identifying obvious data entry mistakes and in identifying outliers and depicting spread of the distribution. This is a good tool used when the sample size is small.

b) Fig. 1 is an individual value plot for the 'Number of cycles to failure' data. It clearly identifies that,

1) the 4th observation 540 473 is much smaller compared to the rest, and,

2) sample variation is increased due to the 4th observation, namely 540 473.

c) Hence we may treat the life 540 473 as an outlier.

5.4.2 Non-Parametric Approach: Box Plot Method:

5.4.2.1 This method is a thumb rule method used to identify the outliers. The first and third quartiles are calculated for the sample. IQR is calculated as the difference between Q3 and Q1 (IQR = Q3–Q1).

a) Arrange data in ascending order:

Cycles Completed
540 473
857 316
889 563
929 103
934 038
938 773

Confirmed Outlier

b) Compute

First quartile: Q1 = 857 316,

Third quartile: Q3 = 934 038

Inter quartile range: IQR= Q3-Q1 = 76 722

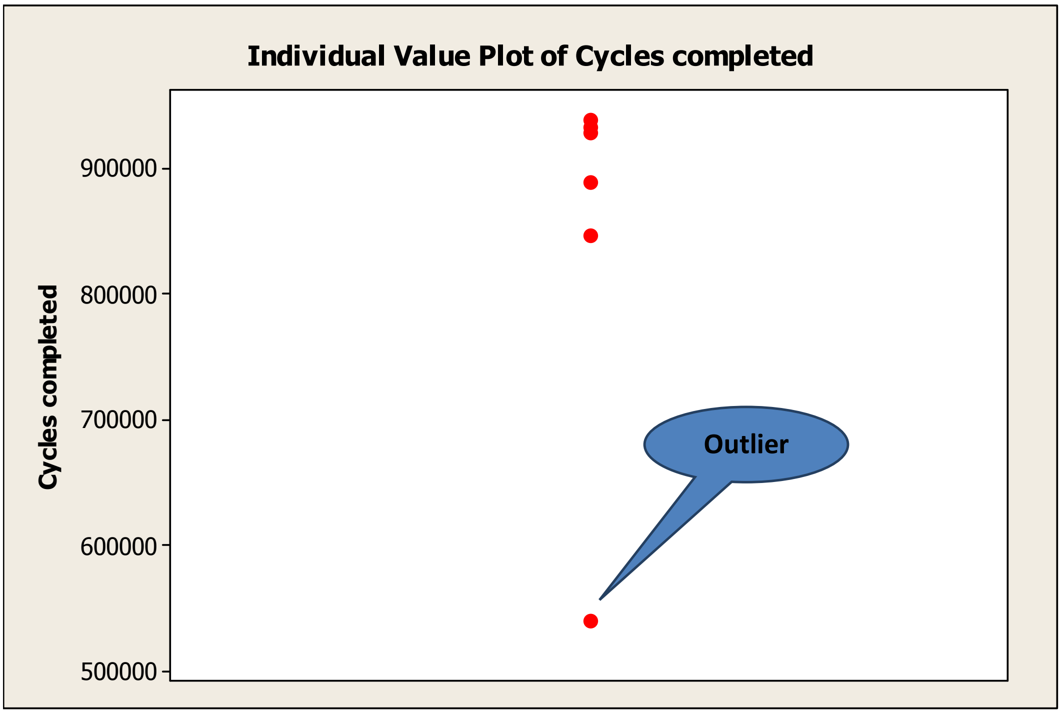

FIG. 1 AN INDIVIDUAL VALUE PLOT FOR THE "NUMBER OF CYCLES TO FAILURE"

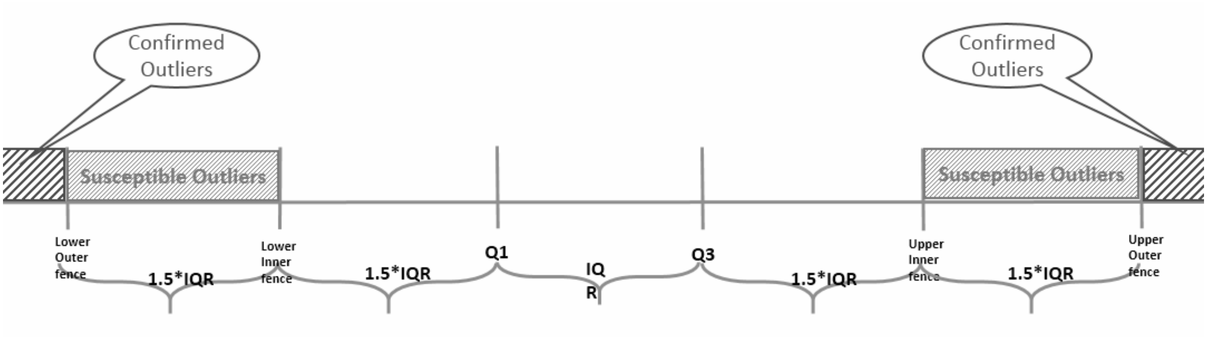

FIG. 2 A SCHEMATIC DIAGRAM REPRESENTING THE FENCES

5.4.2.2 The lower inner and outer fences are calculated as Q1-1.5*IQR and Q1-3*IQR.

a) Lower inner fence = Q1-1.5(IQR) = 742 233
b) Lower outer fence = Q1-3.0 (IQR) = 627 150

5.4.2.3 Similarly, Upper inner and outer fences are calculated as Q3+1.5*IQR and Q3+3*IQR.

a) Upper inner fence = Q1+1.5 (IQR) = 1 049 121
b) Upper outer fence = Q1+ 3.0 (IQR) = 1 164 204

5.4.2.4 As shown in Fig. 2, any point between the inner and outer fence (lower/upper individually) is called as a susceptible outlier. A point beyond the outer fence is called as a confirmed outlier. Since 540 473 lies outside outer fence it is a confirmed outlier.

5.4.3 *Parametric Method*

5.4.3.1 From engineering experience, the variable 'Number of cycles to failure' is likely to follow Weibull distribution. Parameters of the distribution are unknown. As the data size is small PP Plot (for ref.Minitab-16) is recommended to visualize whether the data is from 'Weibull distribution'.

5.4.3.2 As shown in Fig. 3 (PP plot using all six data points), One point outside 95 percent confidence interval is an indicator that the point outside is an 'outlier with 95 percent of confidence'. Fig. 4 is the PP plot when the 'suspected outlier' is deleted from the data set. Following are few observations:

Sl No. (1)	*Criteria* (2)	*Using all Data Points* (*see Fig. 3*) (3)	*Deleting Suspected Outlier* (*see Fig. 4*) (4)	*Inference* (5)
i)	Confidence Intervals	The 4th observation is out of the confidence interval	All points are within the 95 percent Confidence interval	4th observation is an outlier
ii)	Average	The average (MTBF) is 839 861	The average (MTBF) is 908 223	The presence of outlier reduces the MTBF (in this case)
iii)	Standard deviation	The standard deviation is 137 264	The standard deviation is 31 513	Presence of outlier Increases variation
iv)	A D statistics	The value of AD statistic is 3.051	The value of AD statistic is 2.774	As the AD statistic is lower, the distribution is a better fit without the outlier

5.5 All the methods listed above are to identify the outliers. Outliers thus identified could be from different distributions and may be due to different engineering causes. Hence, additional engineering causal analysis is required to understand the reasons behind the low life of 4th observation. Addressing the engineering causes identified would reduce the reappearance of outliers in the tests.

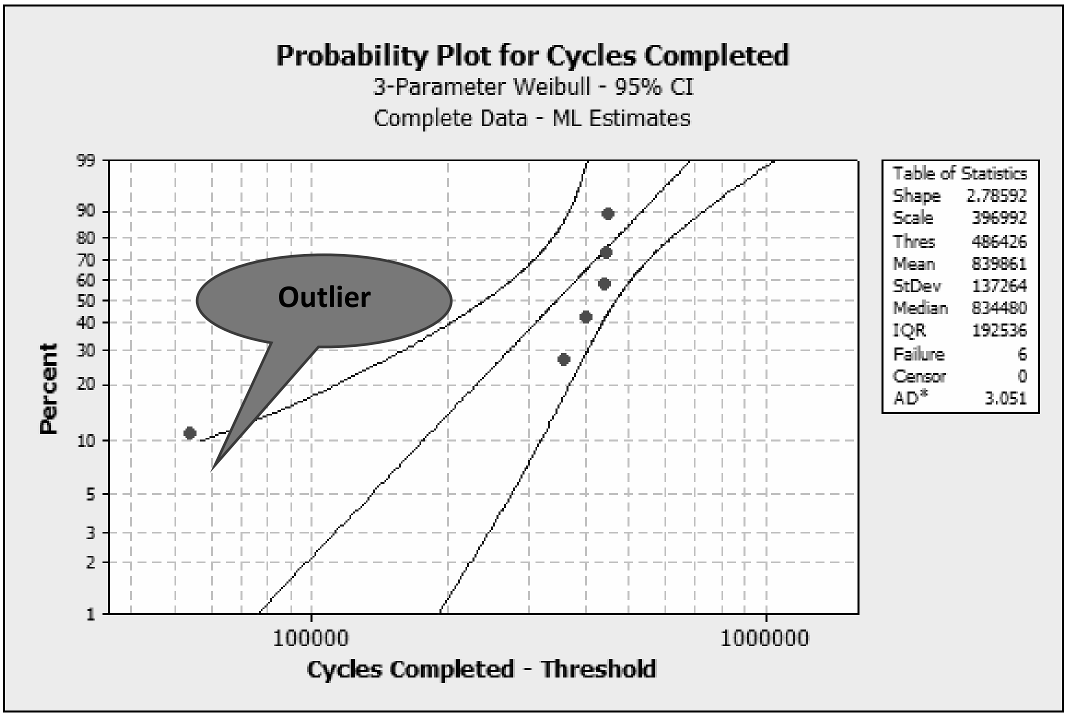

Table of Statistics	
Shape	2.78592
Scale	396992
Thres	486426
Mean	839861
StDev	137264
Median	834480
IQR	192536
Failure	6
Censor	0
AD*	3.051

FIG. 3 PP PLOT USING ALL SIX DATA POINTS. ONE POINT OUTSIDE 95 PERCENT CONFIDENCE INTERVAL IS AN INDICATOR THAT THE POINT OUTSIDE IS AN "OUTLIER"

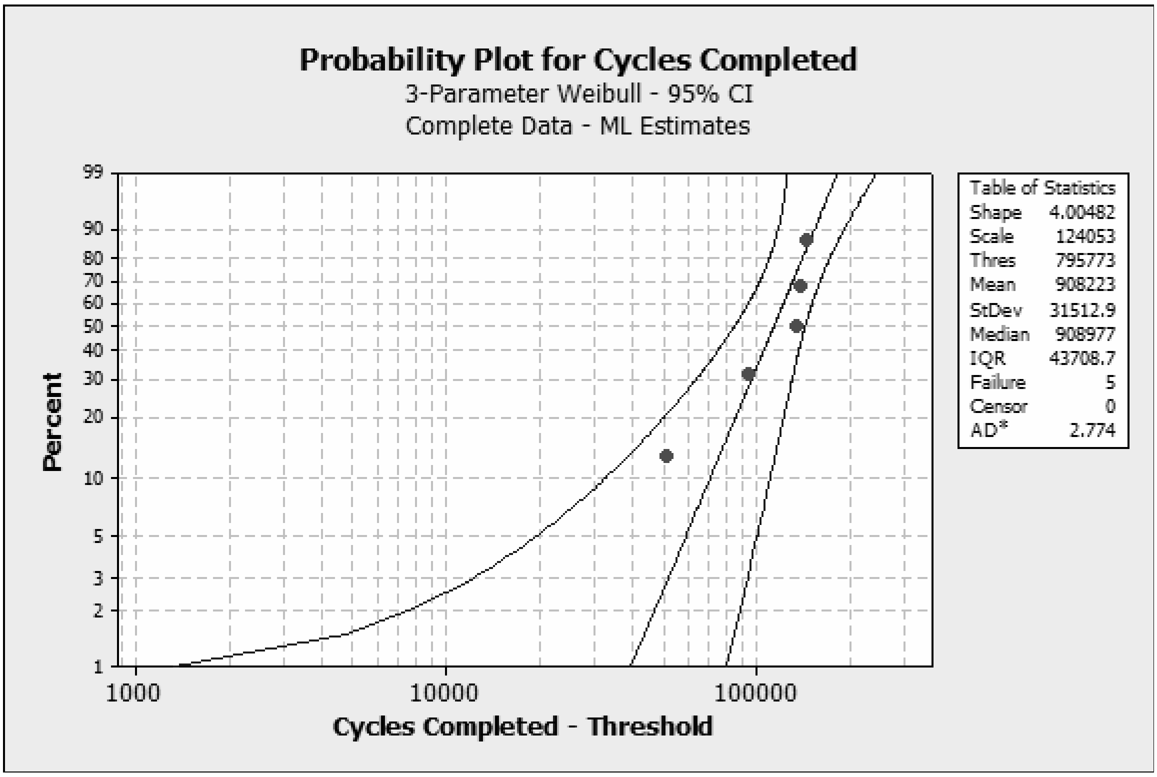

Table of Statistics	
Shape	4.00482
Scale	124053
Thres	795773
Mean	908223
StDev	31512.9
Median	908977
IQR	43708.7
Failure	5
Censor	0
AD*	2.774

FIG. 4 PP PLOT WHEN THE SUSPECTED OUTLIER IS DELETED FROM THE DATA SET

Bureau of Indian Standards

BIS is a statutory institution established under the *Bureau of Indian Standards Act*, 1986 to promote harmonious development of the activities of standardization, marking and quality certification of goods and attending to connected matters in the country.

Copyright

BIS has the copyright of all its publications. No part of these publications may be reproduced in any form without the prior permission in writing of BIS. This does not preclude the free use, in the course of implementing the standard, of necessary details, such as symbols and sizes, type or grade designations. Enquiries relating to copyright be addressed to the Director (Publications), BIS.

Review of Indian Standards

Amendments are issued to standards as the need arises on the basis of comments. Standards are also reviewed periodically; a standard along with amendments is reaffirmed when such review indicates that no changes are needed; if the review indicates that changes are needed, it is taken up for revision. Users of Indian Standards should ascertain that they are in possession of the latest amendments or edition by referring to the latest issue of 'BIS Catalogue' and 'Standards : Monthly Additions'.

This Indian Standard has been developed from Doc No.: MSD 03 (0587).

Amendments Issued Since Publication

Amend No.	Date of Issue	Text Affected

BUREAU OF INDIAN STANDARDS

Headquarters:

Manak Bhavan, 9 Bahadur Shah Zafar Marg, New Delhi 110002
Telephones : 2323 0131, 2323 3375, 2323 9402 *Website*: www.bis.org.in

Regional Offices:		*Telephones*
Central	: Manak Bhavan, 9 Bahadur Shah Zafar Marg NEW DELHI 110002	2323 7617 2323 3841
Eastern	: 1/14 C.I.T. Scheme VII M, V. I. P. Road, Kankurgachi KOLKATA 700054	2337 8499, 2337 8561 2337 8626, 2337 9120
Northern	: Plot No. 4-A, Sector 27-B, Madhya Marg, CHANDIGARH 160019	26 50206 265 0290
Southern	: C.I.T. Campus, IV Cross Road, CHENNAI 600113	2254 1216, 2254 1442 2254 2519, 2254 2315
Western	: Manakalaya, E9 MIDC, Marol, Andheri (East) MUMBAI 400093	2832 9295, 2832 7858 2832 7891, 2832 7892

Branches: AHMEDABAD. BENGALURU. BHOPAL. BHUBANESHWAR. COIMBATORE. DEHRADUN. DURGAPUR. FARIDABAD. GHAZIABAD. GUWAHATI. HYDERABAD. JAIPUR. JAMMU. JAMSHEDPUR. KOCHI. LUCKNOW. NAGPUR. PARWANOO. PATNA. PUNE. RAIPUR. RAJKOT. VISAKHAPATNAM.

Published by BIS, New Delhi